Topics in Natural Language Processing

Shay Cohen

Institute for Language, Cognition and Computation

University of Edinburgh

Lecture 2

Administrativia

Reminder: the requirements for the class are presentations, assignment, brief paper responses and an essay.

- Different topics (and papers) are available online
- They will be of different difficulty levels
- Example topics: topic models, language modeling, parsing, semantics, neural networks (your own topic?)
- Choose whatever level of difficulty you feel comfortable with, so that: (a) your presentation is clear; (b) your brief paper response is informative; (c) the essay goes into details about the topic.

Administrativia

- Presentations start on the week of 8/2
- Please send in the topics to me by Friday next week at 5pm (22/1)
- I will follow-up with an email by some time tomorrow

Last Class

- What is learning?
- What is a statistical model?
- Basic refresher about probability

Θ parameter

$$\underbrace{\{ p(w \mid \theta) \mid \theta \in \Theta \}}_{\text{model}}$$

$p(w \mid \theta)$ is a unigram model

$$\Theta = \{ \theta \mid \theta_i \geq 0, \ \sum \theta_i = 1 \}$$

probability simplex

$$\Omega = \{ w_1, \ldots, w_n \} \qquad p(w_i \mid \theta) = \theta_i$$

Last class: reminder

Probability distributions, random variables, parametrisation

$p(w) \geq 0$ $\quad \sum_w p(w) = 1$ $\quad w \in \Omega$

$X : \Omega \to A$

$Y : \Omega \to B$

$$p(X=x, Y=y) = \sum_{\substack{w \\ X(w)=x \\ Y(w)=y}} p(w)$$

$$X = \begin{cases} 1 & \text{if } w \text{ ends in ion} \\ 0 & \text{o/w} \end{cases}$$

$$Y = \begin{cases} 1 & \text{if } w \text{ begins in } s \\ 0 & \text{o/w} \end{cases}$$

$$p(X=1, Y=1) = \sum_w p(w)$$

station

stipulation

$$p(X=x) = \sum_y p(X=x, Y=y)$$

$$p(Y=y) = \sum_x p(X=x, Y=y)$$

$$p(X=x \mid Y=y) = \frac{p(X=x, Y=y)}{p(Y=y)}$$

Today

- What does statistical learning do?
  - Induce a model from data
  - Models tell us how data is generated
  - Learning does the “opposite”

- Two different paradigms to Statistics: frequentist and Bayesian

Approach 1: frequentist Statistics

- We need an objective function $f(\theta, w_1, \ldots, w_n)$
- The higher the value of f is, the better it predicts the training data

$D = \{w_1, \ldots, w_n\}$

$D \longrightarrow \Theta$

estimation

$$\theta^* = \arg\max_{\theta \in \Theta} f(\theta, w_1, \ldots, w_n)$$

Choice of f: likelihood

$f(\theta, w_1, \ldots, w_n)$ is a real-valued function

$$f(\theta, w_1, \ldots, w_n) = p(w_1 \ldots w_n \mid \theta) = \prod_{i=1}^{n} p(w_i \mid \theta)$$

w_i are independent

Log-likelihood

$$f(\theta, w_1 \dots w_n) = p(w_1 \dots w_n | \theta) = \prod_{i=1}^{n} p(w_i | \theta)$$

$a \geq b$

$\Updownarrow$

$\log a \geq \log b$

We assume w_i are independent

$\log a \cdot b = \log a + \log b$

$$\theta^* = \arg\max_{\theta} \prod_{i=1}^{n} p(w_i | \theta)$$

maximizing likelihood

$$L(w_1 \dots w_n) = f(w_1 \dots w_n)$$

$$\theta^* = \arg\max_{\theta} \log\left(\prod_{i=1}^{n} p(w_i | \theta)\right) = \arg\max_{\theta} \sum_{i=1}^{n} \log p(w_i | \theta)$$

Next step

Estimation: maximisation of L. The result is the "best" θ that fits to the data *according to the objective function* L

$$\theta^* = \operatorname*{argmax}_{\theta \in \Theta} \underbrace{\frac{1}{n} \sum_{i=1}^{n} \log p(w_i \mid \theta)}_{\text{average log-likelihood}}$$

Pre-historic languages

Imagine a language with two words: “argh” and “blah”

Pre-historic languages

What is Ω?

$\Omega = \{ \text{argh}, \text{blah} \}$

What is Θ?

$\Theta = [0, 1]$

θ is the prob. of argh

$1 - \theta$ is the prob of blah

What is the training data?

$w_1 = a \quad w_2 = b \quad w_3, \ldots$

a b a a a b . . . .

Pre-historic languages

What is the likelihood objective function?

$$p(w_i|\theta) = \begin{cases} \theta & \text{if } w_i = arsh \\ 1-\theta & \text{if } w_i = blah \end{cases}$$

$$I(w) = \begin{cases} 1 & \text{if } w_i = arsh \\ 0 & \text{if } w_i = blah \end{cases}$$

$$p(w_i|\theta) = \theta^{I(w_i = arsh)} (1-\theta)^{I(w_i = blah)}$$

$$\log ab = \log a + \log b$$

$$\log a^x = x \log a$$

What is the log-likelihood objective?

$$\log p(w_i|\theta) = I(w_i = arsh)\log\theta + I(w_i = blah)\log(1-\theta)$$

$$L(w_1 \ldots w_n|\theta) = \sum_{i=1}^{n} \log p(w_i|\theta) = \sum_{i=1}^{n} I(w_i = arsh)\log\theta + (1 - I(w_i = arsh))\log(1-\theta)$$

$$= \underbrace{\left(\sum I(w_i = arsh)\right)}_{a}\log\theta + \underbrace{\left(\sum_{i=1}^{n} 1 - I(w_i)\right)}_{b}\log(1-\theta) =$$

$$= a\log\theta + b\log(1-\theta)$$

Pre-historic languages

Log-likelihood: $L(\theta, w_1, \ldots, w_n) = a \log \theta + b \log(1-\theta)$

$(\log \theta)' = \frac{1}{\theta}$

The maximisation problem: $\theta^* = \arg\max_\theta L(\theta, w_1, \ldots, w_n)$

How to maximise this?

$\frac{\partial L}{\partial \theta} = \frac{a}{\theta} + -\left(\frac{1}{1-\theta}\right) \times b = \frac{a}{\theta} - \frac{b}{1-\theta} = 0$

$a(1-\theta) - b\theta = 0$ $\quad a - a\theta - b\theta = 0$

$\theta^* = \frac{a}{a+b}$

~~max~~ MLE

$1 - \theta^* = \frac{b}{a+b}$

Maximisation of log-likelihood

How to maximise the log-likelihood?

Principle of maximum likelihood estimation

- Objective function: log-likelihood (or likelihood)
- Estimation: maximise the log-likelihood with respect to the set of parameters

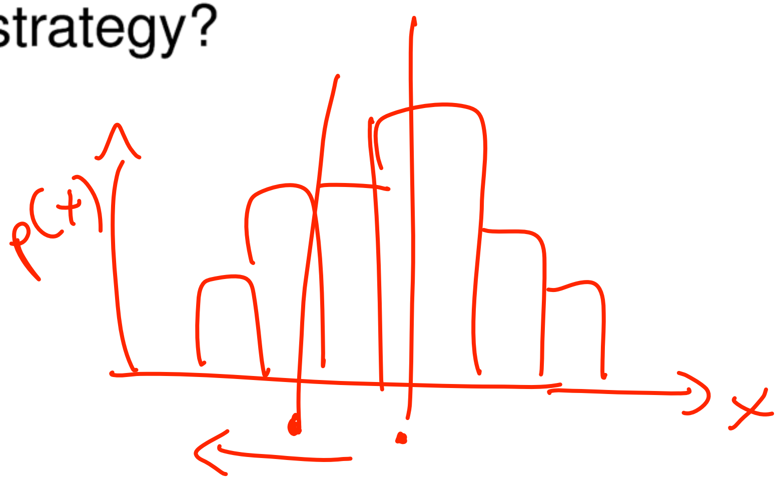
A guessing game

I choose a random number between 1 and 20. You need to guess it, and each time you make a guess I tell you whether your guess is higher or lower than my number. What is your strategy to guess the number as quickly as possible?

A guessing game

I choose a random number between 1 and 20. You need to guess it, and each time you make a guess I tell you whether your guess is higher or lower than my number. What is your strategy to guess the number as quickly as possible?

I choose a random number x between 1 and 20 **from a distribution** $p(x)$. You know p and need to guess the number. What is your strategy?



What does log-probability mean?

Let p be a probability distribution over Ω. What is $-\log_2 p(x)$?

$|code(x)| = -\log_2 p(x)$

code(x) = sequence of 0's and 1's telling whether we make the choice "lower" or "higher"

$E[|code|] = \sum_{\omega} p(\omega)\,|code(\omega)| =$

$= -\sum p(\omega) \log_2 p(\omega)$ ← "entropy"

Another view of maximum likelihood estimation

What is the "empirical distribution?"

$$\tilde{p}(w) = \frac{\text{count}(w \text{ in data})}{n}$$

Rewriting the objective function $L(\theta, w_1, \ldots, w_n)$

$$L(\theta, w_1 \ldots w_n) = \frac{1}{n} \sum_{i=1}^{n} \log p(w_i \mid \theta)$$

$$= \sum_{w \in \Omega} \tilde{p}(w) \log p(w \mid \theta)$$

cross-entropy $(\tilde{p}, p)$

$$\theta^* = \arg\min_{\theta} - \sum_{w} \tilde{p}(w) \log p(w \mid \theta)$$

Cross-entropy

What is the definition of cross-entropy?

Likelihood maximisation

By doing maximum likelihood maximisation we:

- Choose the parameters that make the data most probable,

  or, from an information-theoretic perspective:

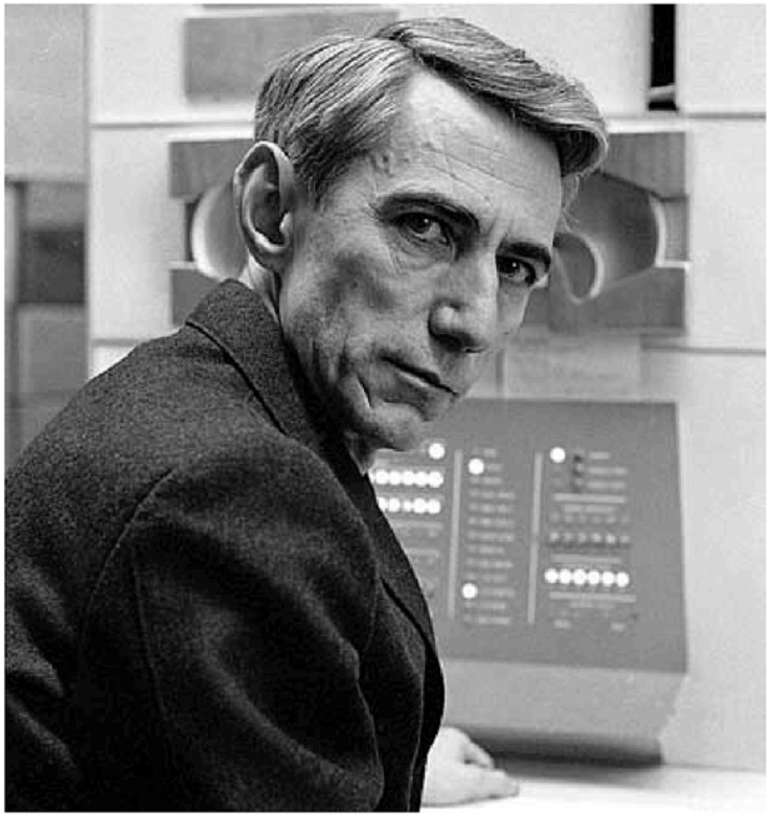
- Choose the parameters that make the encoding of the data most succinct (bit-wise),

  in other words, we

- Minimize the cross-entropy between the empirical distribution and the model we choose.

A bit of history

One of the earliest experiments with statistical analysis of language – measuring entropy of English



2-3 bits are required for English

Approach 2: the Bayesian approach

- History: 1700s. Seminal ideas due to Thomas Bayes and Pierre-Simon Laplace

- A lot has changed since then...

Next class

- The core ideas in Bayesian inference
- Structure in NLP - what type of computational structures are used and how?